Where standard errors come from: some variance algebra

Statistics for Social Scientists II

Bur, GJM

2024-09-14

1 Expectations

At higher levels of math, we talk about "expected values" and "expectations". These are essentially just averages at the population level. Let $\mathbb{P}[Y = y]$ be the probability that a random variable Y takes on a particular outcome y; e.g., if $Y = \text{income}, \mathbb{P}[Y = y]$ means the probability that income takes on this or that particular value.

Then, the expectation of a random variable Y is …

$$\mathbb{E}[Y] := \sum_{y} \mathbb{P}[Y = y] \cdot y$$

For example, suppose that we have a finite population of 100 people, and 40 have no children, 20 have one, 30 have two, and 10 have three children.

$$\begin{aligned}
\mathbb{E}[\text{children}] &= \sum_{y=0}^{3} \mathbb{P}[\text{children} = y] \cdot y \\
&= 0.4 \cdot 0 + 0.2 \cdot 1 + 0.3 \cdot 2 + 0.1 \cdot 3 \\
&= 1.1
\end{aligned}$$

Notice that this is just a shortcut for calculating the average. We could rewrite our last equation like this:

$$\begin{aligned}
\mathbb{E}[\text{children}] &= \sum_{y=0}^{3} \mathbb{P}[\text{children} = y] \cdot y \\
&= \frac{40}{100} \cdot 0 + \frac{20}{100} \cdot 1 + \frac{30}{100} \cdot 2 + \frac{10}{100} \cdot 3 \\
&= \frac{40 \cdot 0 + 20 \cdot 1 + 30 \cdot 2 + 10 \cdot 3}{100}
\end{aligned}$$

This is just equivalent to our formula for the simple mean in a finite population, $\frac{1}{N}\sum_{j=1}^{N} Y_j$—we just have a lot of repeated values (e.g. 30 who had two kids), so we can simply write $30 \cdot 2$. The expectation formula is handy when we don't have a finite population, e.g. if we're considering the abstract possibility that the next child born in the US will be male or female. It also is of course simpler in notational terms.

Note that the three key summation properties—constants factor out, the expectation of a sum of multiple variables is the sum of their expectations, and summing a constant k yields nk—apply to expectations. Here are those properties written out and justified…

$\sum_{j=1}^{n} ky_j = k\sum_{j=1}^{n} y_j$	This is just the distributive property
$\sum_{j=1}^{n} y_j + x_j = \sum_{j=1}^{n} y_j + \sum_{j=1}^{n} x_j$	This is just the associative property
$\sum_{j=1}^{n} k = nk$	Imagine the argument of the sum is actually $0_j + k$ if the lack of an index on k trips

Therefore… (one proof below uses double summations; don't worry too much about these—they basically mean "fix a value for one summation, say $y = 1$, and then run through all $x = 1, 2, 3...$; then, repeat for all values of $y = 2, 3, ...$").

$$
\begin{aligned}
\mathbb{E}[kY] &= \sum_y \mathbb{P}[Y=y]\cdot ky \\
&= k\mathbb{E}[Y] \\
\mathbb{E}[X+Y] &= \sum_x \sum_y \{\mathbb{P}[Y=y\cap X=x]\cdot(x+y)\} \\
&= \sum_x \sum_y \{\mathbb{P}[Y=y\cap X=x]x + \mathbb{P}[Y=y\cap X=x]x\} \\
&= \sum_x \sum_y \mathbb{P}[Y=y\cap X=x]x + \sum_x \sum_y \mathbb{P}[Y=y\cap X=x]y \\
&= \sum_x \mathbb{P}[X=x]x + \sum_y \sum_x \mathbb{P}[Y=y\cap X=x]y \\
&= \sum_x \mathbb{P}[X=x]x + \sum_y \mathbb{P}[Y=y]y \\
&= \mathbb{E}[X] + \mathbb{E}[Y] \\
\mathbb{E}[k] &= \sum_y \mathbb{P}[Y=y]k \\
&= k\sum_y \mathbb{P}[Y=y] \\
&= k
\end{aligned}
$$

Variance as expectation

The variance can be written as an expectation; remember that it's essentially just the mean squared deviation, $\sigma_Y^2 = \frac{1}{N}\sum_{j=1}^{N}(Y_j - \mu_Y)^2$.

$$
\mathbb{V}[Y] := \mathbb{E}[(Y-\mu_Y)^2] = \sum_y \mathbb{P}[Y=y]\cdot(y-\mu_Y)^2
$$

1.1 A useful decomposition

Let's work out the variance of a sample mean across samples. First, we find the variance of a random variable W which is a sum of random variables, $W = \sum_{k=1}^{p} Y_k$.

$$
\begin{aligned}
\mathbb{V}[W] &= \mathbb{E}\left\{\sum_{k=1}^{p} Y_k - \sum_{k=1}^{p} \mu_{Yk}\right\}^2 \\
&= \mathbb{E}\left\{\sum_{k=1}^{p} (Y_k - \mu_{Yk})\right\}^2 \\
&= \mathbb{E}\left\{\sum_{k=1}^{p} (Y_k - \mu_{Yk})^2\right\} + \mathbb{E}\left\{\sum_{i=1}^{p}\sum_{k=1}^{p} (Y_k - \mu_{Yk})(Y_i - \mu_{Yi})\right\} \\
&= \sum_{k=1}^{p} \mathbb{E}\left\{(Y_k - \mu_{Yk})^2\right\} + \sum_{i=1}^{p}\sum_{k=1}^{p} \mathbb{E}(Y_k - \mu_{Yk})(Y_i - \mu_{Yi})] \\
&= \sum_{k=1}^{p} \sigma_k^2 + \sum_{i=1}^{p}\sum_{k=1}^{p} \mathbb{COV}[Y_i, Y_k]
\end{aligned}
$$

This is called the Bienaymé decomposition, and it is extremely useful.

2 Standard errors of means and proportions

We can now obtain the standard errors of means, proportions, and differences in means easily. Each person in our sample is, before they are drawn (i.e., considering the distribution across all possible samples), a random variable.

2.1 Standard error of a single mean

For example, the standard error of a single mean is the variance of a sum divided by a constant n. Since our variables are *independent*—each person's income is observed with a probability that does not change based on whether someone else is observed—they have no covariance with other variables, so that term drops out. *And*, since we assume either *replacement* (unrealistic) or a large population where removing one person doesn't really change the population distribution (realistic), each random variable has the same variance. Together, these two common situations mean that we have *independent, identically-distributed* random variables (IID).

So, we simply write…

$$
\mathbb{V}[\bar{y}] = \mathbb{V}\left\{\frac{1}{n}\sum_{j=1}^{n} y_j\right\}
$$

Now, constants factor out of variances as squares: $\mathbb{E}[(kY - k\mu_Y)^2] = \mathbb{E}[k^2(Y - \mu_Y)^2] = k^2\mathbb{E}[k^2(Y - \mu_Y)^2] = k^2\mathbb{V}[Y]$. So, factor out $\frac{1}{n^2}$ and we have…

$$
\begin{aligned}
\mathbb{V}[\bar{y}] &= \frac{1}{n^2}\mathbb{V}\left\{\sum_{j=1}^{n} y_j\right\} \\
&= \frac{1}{n^2} n\sigma^2 \\
&= \frac{1}{n}\sigma^2 \\
SE &= \frac{1}{\sqrt{n}}\sigma
\end{aligned}
$$

And now we have our famous formula. In the jump from the first to the second line, we used the fact that summing the same variance σ^2 up n times just gives $n\sigma^2$.

2.2 Standard error of a single proportion

We can just apply our previous result from above since a proportion p is just a mean:

$$\mathbb{E}[D] = p = \mathbb{P}[D = 1]$$

for a dummy variable D (you can prove that this is true using the expectation formula given above; it is short).

We just need one additional shortcut, which is to get the variance of that dummy in a simple format, which we'll do with the König-Huygens formula, also extremely useful. Recall that in general $\mathbb{E}[Y] = \mu_Y$; the difference is basically the same difference between the summation formula for the sample mean and the simple symbol $\bar{y}$. I use both below interchangeably.

2.3 The König-Huygens formula

$$
\begin{aligned}
\mathbb{V}[Y] &= \mathbb{E}[(Y - \mu_Y)^2] \\
&= \mathbb{E}[Y^2] - 2\mu_Y\mathbb{E}[Y] + \mathbb{E}[Y^2] \\
&= \mathbb{E}[Y^2] - 2\mu_Y \cdot \mu_Y + \mu_Y^2 \\
&= \mathbb{E}[Y^2] - \mu_Y^2 \\
&= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2
\end{aligned}
$$

This formula very often simplifies proofs and identities. Here, it really simplifies the formula for the variance of a dummy.

$$
\begin{aligned}
\mathbb{V}[D] &= \mathbb{E}[D^2] - \mathbb{E}[D]^2 \\
&= \sum_{d=0}^{1} \mathbb{P}[D = d] \cdot d^2 - p^2 && \text{probability is still just } D = d \\
&= \mathbb{P}[D = 0] \cdot 0^2 + \mathbb{P}[D = 1] \cdot 1^2 - p^2 \\
&= p - p^2 \\
&= p(1 - p)
\end{aligned}
$$

Since a sample proportion is a mean, we just combine this with our previous result:

$$
\begin{aligned}
\mathbb{V}[D] &= p(1 - p) \\
\mathbb{V}[\bar{y}] &= \frac{1}{n}\sigma^2 \\
\mathbb{V}[\hat{p}] &= \frac{1}{n}p(1 - p)
\end{aligned}
$$

2.4 Sampling variance of a difference in sample means

Now, we just the variance of a difference in sample means. This is sort of an interesting case where we use the same formula two different ways: we will use the Bienaymé formula on two different levels here. First, we have a different in means. If we let W be the difference in means $\bar{y}_H - \bar{y}_L$ (higher group less lower group, which is almost always the most convenient way to write the difference), we have…

$$
\begin{aligned}
\mathbb{V}[W] &= \sum_{k=1}^{p} \sigma_k^2 + \sum_{i=1}^{p} \sum_{k=1}^{p} \mathbb{COV}[Y_i, Y_k] \\
&= \sigma_{\bar{y}_H}^2 + \sigma_{\bar{y}_L}^2 + 2\mathbb{COV}[\bar{y}_H, \bar{y}_L]
\end{aligned}
$$

Three things should be mentioned now. First, I cheated a bit and used the formula for a sum, not a difference, but $\mathbb{V}[-Y] = \mathbb{V}[Y]$ by the "constants factor out as squares" rule above, so it makes no difference. Second, we have no covariance here if the groups are selected independently; if they *aren't* (suppose our difference is between husbands and wives), that's fine: it makes the standard error smaller! I show how to find that covariance in the appendix. Third, we already know $\sigma_{\bar{y}_H}^2$ and $\sigma_{\bar{y}_L}^2$: we found them above. So, we simply have…

$$\begin{aligned}\mathbb{V}[W] &= \sigma^2_{\bar{y}_H} + \sigma^2_{\bar{y}_L} \\ &= \frac{\sigma^2_H}{n_1} + \frac{\sigma^2_H}{n_2} \\ SE &= \sqrt{\frac{\sigma^2_H}{n_1} + \frac{\sigma^2_H}{n_2}}\end{aligned}$$

This formula might *look* like a crude mistake, an attempt to average the two variances, but that is an unfortunate coincidence. It is just the result of variance algebra.

Note that if we assume *equal* variance, we simply have…

$$\begin{aligned}SE &= \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \\ &= \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\end{aligned}$$

A crucial point in all of this is that while it is *true* that a difference in sample means is only *approximately t-distributed*, the variance of a difference of two random variables is *exact* and *does not depend on the distributional form*. You might notice that *throughout* these notes, we have not had to assume anything about the overall *shape* of the sampling distribution, even though we could still get its *parameters*.

3 Estimation

Fortunately, we are on solid statistical ground if, in order to estimate these quantities, we just plug in the sample estimates. This is a procedure known as the "method of moments" that works in most simple cases. So, just plug in your sample estimate of any unknown above to get an unbiased estimator of the sampling variance.

4 A demonstration using Python

```
import numpy as np
import pandas as pd
import random

# Let's set the parameters for our population as being the following...

sigma_b, mu_b, n_b = 3, 6, 28
sigma_g, mu_g, n_g = 2, 7.2, 34
samples = [] # We'll fill this in with samples soon
rng = np.random.default_rng() # Set the instance of the random number generator
```

```
random.seed(4301963) # Set the seed (best to pick large numbers; I just use a birthday)
for samp in range(100000):
    # We'll pull 100,000 samples--enough to approximate the true sampling distro
    samp1 = rng.normal(mu_b, sigma_b, n_b)
        # I pull from a Normal distro
    samp2 = rng.normal(mu_g, sigma_g, n_g)
    diff = samp2.mean() - samp1.mean()
        # Find the sample difference
    samples.append(diff)
        # And now we append it to the data set of sampling differences
        # Essentially, that's a sampling distribution

true_var = np.round(np.var(samples), 5) # The empirical variance
theoretical_var_uneq = \
    np.round(sigma_b**2/n_b + sigma_g**2/n_g, 5) # The one using the unequal formula
theoretical_var_eq = \
    np.round(((n_g-1)*(sigma_g**2)  + (n_b-1)*(sigma_b**2))/(n_b*n_g), 5) # Using the equal

print(f"The true variance of the sample differences in reality is {true_var}")
print(f"The theoretical variance using the unequal formula is {theoretical_var_uneq}")
print(f"The theoretical variance using the pooled formula is {theoretical_var_eq}")
```

```
The true variance of the sample differences in reality is 0.43963
The theoretical variance using the unequal formula is 0.43908
The theoretical variance using the pooled formula is 0.39391
```

5 Appendix: covariance of sample means

Let's start with the covariance of two sums. In what follows, notation such as Y_i^H means the ith random variable taken from the distribution Y^H, where H means one outcome of a grouping variable. E.g., Y_4^{male} would be "the income (education, height, etc.) of the fourth male we draw". We use this notation to keep track of the difference between identically-distributed variables that represent the same conceptual thing (e.g. Y_4^{male} vs. Y_5^{male}) and the same conceptual thing but from different subgroups' distributions (e.g. Y_4^{male} vs. Y_4^{female}).

$$
\begin{aligned}
\mathbb{COV}\left\{\sum_{i=1}^{n} Y_i^H, \sum_{j=1}^{m} Y_j^L\right\} &= \mathbb{E}\left\{[\sum_{i=1}^{m}(Y_i^H - \mu_H)][\sum_{j=1}^{n}(Y_j^L - \mu_L)]\right\} \\
&= \mathbb{E}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n}[(Y_i^H - \mu_H)][(Y_j^L - \mu_L)]\right\} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}[(Y_i^H - \mu_H)][(Y_j^L - \mu_L)] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{COV}[Y_i^H, Y_j^L]
\end{aligned}
$$

Now, we just write our means as normed sums:

$$\mathbb{COV}\left\{\frac{1}{n}\sum_{i=1}^{m} Y_i^H, \frac{1}{m}\sum_{j=1}^{n} Y_j^L\right\} = \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{COV}[Y_i^H, Y_j^L]$$

In social sciences, if two samples are related, they are sometimes *matched pairs*, as noted above. Then, their relation is the following, since *only* people within a pair are related—not across pairs (also note that now $n = m$ logically).

$$\begin{aligned}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{COV}[Y_i^H, Y_j^L] &= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{COV}[Y_k^H, Y_k^L] \\ &= \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{COV}[Y_k^H, Y_k^L] \\ &= \frac{1}{n^2}n\mathbb{COV}[Y^H, Y^L] \\ &= \frac{1}{n}\mathbb{COV}[Y^H, Y^L]\end{aligned}$$

The final steps in the proof follow from the fact that we have IID random variables, just as above.